DOCUMENT RESUME

ED 371 019 TM 021 650

AUTHOR Wang, Yu-Chung Lawrence
TITLE Robustness of Unidimensional IRT Calibration in the Presence of Essential Dimensionality.
PUB DATE Apr 94
NOTE 21p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 4-8, 1994).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Correlation; Educational Assessment; Estimation (Mathematics); *Item Response Theory; *Robustness (Statistics); *Sample Size; *Selection; Test Format; *Test Items
IDENTIFIERS *Calibration; DIMTEST (Computer Program); Invariance; Item Calibration; *Unidimensionality (Tests)

ABSTRACT

The first purpose of this study was to investigate the stability of two essential dimensionality measures across 10 random samples within a particular assessment item (AT1) selection. Other purposes were to investigate the discrepancy of the essential unidimensionality estimates for a test across different AT1 selections and sample sizes and to investigate the validity of replacing the item response theory (IRT) unidimensionality assumption with the essential unidimensionality assumption using the existence of the invariance property of item parameters as a criterion. Results indicate that the stability of two essential unidimensionality measures is low for some tests across 10 random samples, but the correlation is high within the same sample. The essential dimensionality results for four tests across four different AT1 assignments were also different, indicating that the essential dimensionality estimate for a test is related to the characteristics of AT1 items. It was found that reducing sample size or reducing the number of test items and AT1 items does not assure unidimensionality. Relationships between the existence of the item invariance property and the essentially unidimensional item calibrations are low across test forms and mathematics areas. A further study of the criteria of AT1 items is needed to enhance the validity of replacing the IRT unidimensionality assumption by the essential unidimensionality assumption. Fourteen tables present the data. (Contains 13 references.) (Author/SLD)

Robustness of Unidimensional IRT Calibration in the Presence of Essential Dimensionality

Yu-Chung Lawrence Wang

University of Southern California

Paper presented at annual meeting of the American Educational Research Association, April 1994, New Orleans, LA.

The author wish to thank Ian Westbury for providing access to the SIMS data. This research was partially supported by the Educational Policy Fellowship during the first year of the study and by Chiang Ching-Kuo Research Scholarship in the second year of the study.

Robustness of Unidimensional IRT Calibration in the Presence of Essential Dimensionality

Abstract

Stout (1987, 1990) has argued that the essential dimensionality assumption is a valid substitute for Lord's unidimensionality assumption (Lord, 1980). The first purpose of this study was to investigate the stability of two essential dimensionality measures across ten random samples within a particular AT1 selection. The second was to investigate the discrepancy of the essential dimensionality estimates for a test across different AT1 selections and sample sizes. Finally, the third purpose was to investigate the validity of replacing the IRT unidimensionality assumption with the essential unidimensionality assumption using the existence of the invariance property of item parameters as a criterion. The results of this study indicated that the stability of two essential dimensionality measures was low for some tests across ten random samples. The correlation between two different essential dimensionality measures was high within the same sample. The essential dimensionality results for four tests across four different AT1 assignments was also different. This finding indicates that the essential dimensionality estimate for a test is related to the characteristics of the AT1 items. In the second section of analysis, the effect of reducing the number of examinees and test items was analyzed. It was found that reducing sample size does not provide consistent improvement on the degree of essential unidimensionality. Also, reducing the number of test items and AT1 items did not assure unidimensionality. The characteristics of the AT1 items likely has more influence on essential dimensionality estimates. The validity of replacing the IRT unidimensionality assumption by the essential unidimensionality assumption was assessed in the last section of the analysis using the invariance of item parameters as evidence. It was universally found that the relationships between the existence of the item invariance property and the essentially unidimensional item calibrations (i.e., arithmetic and algebra scale) are low across test forms and mathematic areas. Therefore, a further study on the criteria of AT1 items is needed to enhance the validity of replacing the IRT unidimensionality assumption by the essential unidimensionality assumption.

Item response theory (Lord, 1980) has been widely used in test equating and test bias studies because of its unique invariance property for item parameters and ability estimates. The invariance property of IRT exists only when the assumptions of unidimensionality and local independence are true. The classical unidimensionality assumption however has been criticized as not realistic in a real test (Traub, 1983). Stout (1987, 1990) has provided a weaker and psychologically more meaningful assumption called essential dimensionality, and has argued that the essential dimensionality assumption is a valid substitute for Lord's unidimensionality assumption. The strengths of Stout's essential dimensionality concept are his objective essential dimensionality statistics and the corresponding easy-to-use computer program (DIMTEST, 1992). Stout's contributions, however, encounter four challenges . First of all, the degree of essential dimensionality of a group of target items (or a test) depends heavily on the characteristics of the assessment items (i.e., AT1) that are chosen. For instance, Wang and Hocevar (1994) have found that a group of items in the arithmetic area may be flagged as essentially unidimensional when they are compared to decimal fraction AT1 items, but may be multidimensional when ratio proportion AT1 items are chosen. Secondly, both of Stout's essential dimensionality statistics are subject dependent because the original dichotomous item responses are used in the DIMTEST program. Therefore, two groups of examinees with distinct cognitive skills may interact differently with a particular group of AT1 items and result in two magnitudes of essential dimensionality. Thirdly, the number of items and examinees may also influence the result of an essential dimensionality analysis since it may be easier to have a unidimensional test with fewer items as well as fewer subjects. Finally, there is no clear cutoff score that the user can use to conclude that essential dimensionality is totally equivalent to Lord's unidimensionality assumption. Beyond all these problems, the stability of the two essential dimensionality statistics across random samples is also unknown.

There were three purposes to this study. The first purpose was to investigate the stability of the two essential dimensionality measures across ten random samples within a particular AT1 selection. The second was to investigate the discrepancy of the essential dimensionality estimates for a test across different AT1 selections and sample sizes. Finally, the third purpose was to investigate the validity of replacing the IRT unidimensionality assumption with the essential dimensionality assumption using the existence of the invariance property of item parameters as a criterion.

Methodology

Data.

The data for this study were adapted from the Second International Mathematics Study (SIMS) of the International Association for the Evaluation of Educational Achievement (1985). During 1980 and 1982, the Second IEA International Mathematics Study researchers collected data on mathematics curricula, teaching practices, and achievement from samples of students, teachers, and schools in 20 countries. SIMS was conducted at two levels: (1) Population A in which students were (typically) in the national grade in which the modal age was 13; and (2) Population B where students were taking the most advanced pre-

university mathematics course(s) offered in their school systems. However, only the U.S. population A study of SIMS was considered in the present study. Population A of the U.S. study is defined as the eighth graders in both main-stream and non-public schools. Mentally, physically, emotionally, or learning disabled students who were placed into special education class were not included. Subjects were then selected by using school type, regional standard metropolitan statistical area (SMSA), location, and metropolitan status as stratification variables.

Instruments.

There were two major SIMS study designs: longitudinal and cross-sectional. The U.S. was in the longitudinal study design. For the U.S. instrument, an eighth-grade mathematics achievement test that consisted of a total of 180 items which were selected from the international bank of 196 items divided into a 40-item core subtest and four 35-item "rotated forms" was used. The 40 core items were administered to all examinees, and rotated forms were randomly assigned to students (approximately one- fourth of the students taking each form). In other words, each student was administered the core and one rotated form for a total of 75 of the 180 items in the pool. Every SIMS achievement test covers five major content areas: arithmetic, algebra, geometry, statistics and measurement. There were several subcontent categories under each major content area. For example, there were eight subcontent areas within the arithmetic area: Natural Numbers (001), Common Fractions (002), Decimal Fractions (003), Ratios, Proportions, and Percent (004), Number Theory (005), Power and Exponents (006), Square Roots (008), and Dimensional Analysis (009). Four test forms of eighth grade mathematic tests (form A, B, C, and D) were investigated in the U.S. study and were recoded as test 1 to test 4. Also every subtest was relabeled for this study (see Table 1). The first number of the subtest denotes the resource of the subtest, and the extension number 1, 2, 3, and 4 respectively denotes arithmetic, algebra, geometry, and measurement.

In the stability analysis using ten random sample data subsets, only the arithmetic and algebra tests which had an appropriate number of items for AT1 assignment (i.e., common fractions (002), decimal fractions (003), and ratios (004) for arithmetic, and integers (101), formulas (104), and equations (106) for algebra) were chosen for analysis. Readers who are interested in the complete content of all SIMS items should refer to "Technical Report I" of SIMS (Chang & Ruzicka, 1985).

To examine the effect of sample size on DIMTEST estimates, four different numbers of examinees were chosen (n=1600, 800, 400, 200). First samples were selected from the original data with an arbitrary fixed number of 1600 examinees. This 1600-case dataset was then randomly split into three mutually exclusive subsamples (i.e., 200, 400, 800).

Table 1. Labels of Tests and Subtests in SIMS U.S. Study.

Test Recode	Items	N	Description
1	75	1652	Test A
1.1	28	1652	Subtest of test A (Arithmetic)
1.2	14	1652	Subtest of test A (Algebra)
1.3	17	1652	Subtest of test A (Geometry)
1.4	12	1652	Subtest of test A (Measurement)
2	75	1610	Test B
2.1	28	1610	Subtest of test B (Arithmetic)
2.2	14	1610	Subtest of test B (Algebra)
2.3	17	1610	Subtest of test B (Geometry)
2.4	12	1610	Subtest of test B (Measurement)
3	75	1668	Test C
3.1	27	1668	Subtest of test C (Arithmetic)
3.2	14	1668	Subtest of test C (Algebra)
3.3	16	1668	Subtest of test C (Geometry)
3.4	13	1668	Subtest of test C (Measurement)
4	75	1619	Test D
4.1	27	1619	Subtest of test D (Arithmetic)
4.2	14	1619	Subtest of test D (Algebra)
4.3	16	1619	Subtest of test D (Geometry)
4.4	13	1619	Subtest of test D (Measurement)

To investigate the effects of reducing the number of AT1 item on DIMTEST estimates, two types of subsamples were generated. In substudy one, the first 28 items were chosen from each 75-item test 1, 2, 3, and 4 and the essential dimensionality estimates for these subtests were calculated using 3 randomly selected AT1 items. The purpose of this study was to determine if the previously reported existence of unidimensionality in more content specific tests (see Wang & Hocevar, 1994) is related to a reduction of the number of AT1 items. In the second substudy the effect of reducing the number of AT1 items was investigated more systematically using three different numbers of randomly selected AT1 items --- 12, 8, and 4.

To examinee the invariance property of item parameters in an essentially unidimensional test, arithmetic and algebra items in tests 1, 2, 3, and 4 were examineed using two randomly selected, mutually exclusive, split-half samples with 1/2 of the original sample size of examinees.

Analysis.

There are three sections in the analysis. In the first section, a replication study design was applied to test the reliability of Stout's statistics across random samples. That is, ten sample data sets with approximately one-fourth of the original sample size of examinees (N= 400) were randomly selected from the original data for each test. Furthermore, there were three AT1 selections, (common fractions, decimal

fractions and ratios) one for each of the four arithmetic tests. Along the same lines, there were three algebra AT1 selections (integers, formulas, and equations) for each of the four algebra tests. The essential dimensionality of both arithmetic and algebra for the four different test forms (i.e., test 1, 2, 3, and 4) were calculated ten times using each random sample described earlier. The means, standard deviations and ranges of the ten replicated Stout's essential dimensionality statistics for arithmetic and algebra tests were calculated and used as indices to assess the consistency of DIMTEST across random samples.

To examine the effect of sample size on the essential dimensionality estimates, four mutually exclusive random subsamples from four SIMS original tests with 200, 400, 800, and 1600 cases were investigated in the second section of the analysis. The measurement items in each test were chosen as AT1 items and fixed across the four subsamples. The effect of reducing the number of AT1 items on DIMTEST also was investigated in this section of analysis by matching the number of total items and AT1 items between the general mathematic tests (test 1, 2, 3, 4) and content specific tests (test 1.1, 2.1, 3.1 and 4.1).

The idea of the essential dimensionality statistic is to assess whether there is a dominant trait measured by the test. Stout (1990) suggests that multidimensional item characteristics and abilities are suitable for unidimensional IRT as long as there is a dominant trait. Based on this argument, the item invariance property should fit when unidimensionality is replaced with essential dimensionality. The purpose of the last section of this study was to determine the equivalence between the IRT unidimensionality assumption and essential dimensionality by investigating the relationship between essential unidimensionality and the item invariance property. The item invariance property was then investigated for essentially multidimensional test 1 and the essentially unidimensional arithmetic and algebra test within test 1 (or test 1.1 and 1.2) using Lord's chi-square (1980), and Raju's two exact area measures (1988) to assess invariance.

Three corresponding research questions are: first, are two measures of essential dimensionality stable across samples ? ; second, do smaller sample tests and tests with fewer AT1 items have higher degree of the essential unidimensionality ? ; and third, is the degree of the fit of the item invariance property for a test associated with the degree of the essential unidimensionality of the test ?

Results

Stability of DIMTEST.

Table 2 displays the summary results of means and standard deviations of the two essential dimensionality statistics for four arithmetic tests, 1.1 to 4.1, with three arithmetic subarea AT1 items using ten randomly selected samples. Surprisingly, both of Stout's essential dimensionality T estimates vary across random samples. To illustrate, the range of Stout's T and T' measures is found to be as high as 3.00 and 3.66, respectively. And further, some tests were identified as essentially multidimensional almost as many times as they were identified as essentially unidimensional within the same AT1 situation. For example, test 1.1 was identified as essentially unidimensional only six times out of a total of ten trails

using common fraction (002) AT1 items. This result indicates that the stability of essential dimensionality statistics across random sample is fairly low.

As in a previous study (Wang & Hocevar, 1994), the degree of the essential dimensionality for a test was found to be highly associated with the characteristics of AT1 items. For example, Test 1.1 which was identified as either 60%, 90% and 80% essentially unidimensional depending on whether common fractions, decimal fractions, and ratio AT1 items were used (Refer to Table 2). In other words, the degree of the essential dimensionality of a test depends on the set of AT1 items that was assigned.

In addition, it is noteworthy that the difference between the original and refined T statistic is negligable, except in the case of test 2.1 with common fractions AT1 items. However, Stout's original T score tends to be more reliable due to its lower range and SD across ten estimates than its refined counterpart (T'). The refined essential dimensionality statistic, on the other hand, tends to be more powerful. A detailed comparison study of these two statistics is needed for a more conclusive interpretation for their results.

Furthermore, Table 2 demonstrates that the degree of the essential dimensionality for the four test forms is not the same. That is tests 3.1 and 4.1 generally were found to be more essentially unidimensional than tests 1.1 and 2.1. This discrepancy may be attributed to the effects of test forms. In other words, the discrepancy of the essential dimensionality within four tests may due to different item compositions in the four tests. However, it is important to point out that the four forms were created randomly.

Table 3 presents the results of an identical analysis on four algebra tests using three different AT1 assignments. It is shown that even though the magnitude of the largest range of two Stout's statistics (2.62 and 3.40 in the algebra tests) is about the same as the largest magnitude in the arithmetic subtests displayed earlier, the acceptance rates of the essential dimensionality assumption for the algebra tests are higher and more consistent than the arithmetic tests. This result implicates that the degree of the essential unidimensionality for four algebra tests is higher than their four arithmetic counterparts. However, test 4.2 is the only test that was identified as essentially unidimensional across all three AT1 selections. The degree of consistency of essential unidimensionality estimates across the three AT1 forms is also higher than in the arithmetic counterpart. These findings indicate either SIMS algebra subtests are more essentially unidimensional than the arithmetic counterparts or the U.S. students' cognitive abilities in algebra are more homogeneous than their cognitive abilities in arithmetic.

Again, comparing Stout's essential original dimensionality measure with its refined counterpart, both Table 2 and Table 3 show that Stout's original statistic is more consistent than the refined statistic as indicated by a smaller standard deviation and range. However, as mentioned imediately above, highly consistent unidimensional "flags" were noted on virtually all algebra tests (Table 3) by both T and the refined T' !

Table 2. Means, Standard Deviations, and Acceptance Rates for Essential Dimensionality for Four Arithmetic Test Forms Using Three AT1 Assignments.

Resource Test	AT1	Mean T	SDT	Mean T'	SDT'	RT*	RT'	PT**	PT'
1.1		1.11	0.99	1.35	1.32	2.93	3.53	0.60	0.60
2.1	Com	0.78	0.96	1.00	1.22	3.00	3.66	0.90	0.60
3.1	Frac.	0.51	0.50	0.64	0.59	1 51	1.83	1.00	0.90
4.1		0.38	0.56	0.43	0.65	1.59	1.87	1.00	1.00
1.1		0.73	0.54	0.87	0.67	1.92	2.37	0.90	0.90
2.1	Deci.	-0.31	0.64	-0.33	0.76	1.62	1.92	1.00	1.00
3.1	Frac.	0.06	0.45	0.08	0.54	1.50	1.81	1.00	1.00
4.1		0.57	0.64	0.71	0.81	1.77	2.34	1.00	0.90
1.1		0.74	0.85	0.79	1.01	2.61	3.25	0.90	0.80
2.1	Ratio	1.16	0.47	1.32	0.56	2.77	3.27	0.90	0.70
3.1	Prop.	1.12	0.73	1.26	0.88	2.65	2.99	0.80	0.70
4.1		1.41	0.60	1.59	0.70	2.49	2.92	0.70	0.60

* R denotes the range between maximum and minimum estimates.
** P denotes percent of times the test was flagged as essentially unidimensional.

Table 3. Means, Standard Deviations, and Acceptance Rates for Essential Dimensionality Statistics for Four Algebra Tests Using Three AT1 Assignments.

Resource Test	AT1	Mean T	SDT	Mean T'	SDT'	RT	RT'	PT	PT'
1.2		0.31	0.32	0.34	0.41	0.97	1.28	1.00	1.00
2.2	Integ.	0.91	0.60	1.07	0.75	2.42	2.85	0.90	0.90
3.2		0.49	0.31	0.63	0.40	1.00	1.29	1.00	1.00
4.2		0.46	0.43	0.49	0.50	1.57	1.77	1.00	1.00
1.2		0.26	0.77	0.35	0.97	2.62	3.40	1.00	0.90
2.2	Form	-0.40	0.62	-0.48	0.73	1.79	2.07	1.00	1.00
3.2		-0.47	0.86	-0.56	1.00	2.81	3.27	1.00	1.00
4.2		-0.26	0.56	-0.37	0.75	1.67	2.14	1.00	1.00
1.2		-0.45	0.60	-0.44	0.71	1.68	1.99	1.00	1.00
2.2	Equa.	-0.23	0.52	-0.25	0.61	1.61	1.85	1.00	1.00
3.2		0.01	0.91	0.05	1.10	2.91	3.52	0.90	0.90
4.2		0.62	0.40	0.77	0.49	1.13	1.45	1.00	1.00

Effect of Sample Size and AT1 Size.

The effect of item size on essential dimensionali easures were examined in the following analyses. To examine the effect of sample size on DIMTEST estimates, four different numbers of U.S. examinees were selected for each test. First a sample was selected from original data for general mathematics achievement test 1 with an arbitrary fixed number of examinees, which was 1600. The 1600-case dataset was then randomly split into three mutually exclusive subsamples (i.e., 200, 400, 800). That

is, a dataset was exclusively selected for test 1 with sample size 1600, 800, 400, and 200, as subsamples test 1A, 1B, 1C, and 1D, respectively. Test 2A to 2D are subsamples of test 2, and so on.

Table 4.1 presents the essential dimensionality assessment for the four SIMS tests using the four different sample sizes. Test 1 was flagged as essentially unidimensional in the test 1A and 1B situations, and DIMTEST encountered estimation problems when sample size was reduced to 400 for test 1. The degree of the essential multidimensionality slightly increased when the sample size decreased from 1600 to 800 in test 1. Test 2 was flagged as multidimensional when the sample size was 1600 and 800, but the test was identified as essentially unidimensional when the sample size was reduced to 400. The dimensionality result for test 3, however, shows that the degree of the essential unidimensionality increased (from p=.07 to p=.15) when the sample size was reduced from 1600 to 800, but the p-value decreased to .06 when the sample size was reduced to 400. The effect of the sample size on the essential dimensionality for test 4 has the opposite pattern; that is, the essential dimensionality decreased first when the sample size reduced to 800 and jumped back to the similar level of essential dimensionality for size 1600 when the sample size reduced to 400. This result indicates that smaller sample size may increase as well as decrease the degree of the essential unidimensionality estimate.

The previous results suggests two conslusions: first, changes in sample size affect DIMTEST estimates in an unpredictable, but fairly minor way. That is, changes in the estimates may be totally attributable to normal sample error. Second, it is noteworthy that small sample size cause some convergence problems on some occasions for DIMTEST. Thus, it appears that N=400 might be considered a minimum sample size for DIMTEST.

To investigate whether the previously reported greater degree of essential unidimensionality in the analysis of arithmetic and algebra tests (see Wang & Hocevar, 1994) is due to a smaller number of items in these analyses, a new multidimensional general achievement test was generated by arbitrarily assigning the first twenty-eight items from tests 1 to 4. The essential dimensionality estimates were calculated four times for each new test using four different sets of three randomly selected AT1 items. The goal of this analysis was to determine the effects of reducing the number of total items on DIMTEST estimates. Table 4.2 presents the essential dimensionality results for the 28-item subtests. Only one out of a total of sixteen trials shows the predicted significant result. According to this result, a general achievement test with multidimensional items may be flagged as unidimensional because the number of the AT1 items is small. Moreover, the critical p-values for every trial were not highly consistent. For instance, the four p-values for test 2 fall into the range from .90 to .15. Test 3 was identified as multidimensional (with p-value equals .02) when the first three items in the test were AT1 items, but was flagged as unidimensional (p-value equals .79) when other items were AT1. This discrepancy, again, shows that the actual AT1 characteristics has a significant effect on the essential dimensionality estimates.

Table 4.1 Essential Dimensionality for Four SIMS General Mathematics Achievement Tests and Subsamples Using Measurement AT1 Items.

Test	Sample size	T	P	T'	P'
Subsample of test 1.					
1A	1600	-.11	.54	-.12	.55
1B	800	-.20	.58	-.23	.59
1C	400	Error due to small sample size.			
1D	200	Error due to small sample size.			
Subsample of test 2					
2A	1600	2.63	.00**	3.08	.00**
2B	800	2.08	.02*	2.50	.00**
2C	400	1.10	.13	1.22	.11
2D	200	Error due to small sample size.			
Subsample of test 3					
3A	1600	1.45	.07	1.77	.04*
3B	800	1.04	.15	1.33	.09
3C	400	1.52	.06	2.09	.02*
3D	200	Error due to small sample size.			
Subsample of test 4.					
4A	1600	-.47	.68	-.62	.73
4B	800	.15	.44	.12	.45
4C	400	-.45	.68	-.52	.70
4D	200	Error due to small sample size.			

Table 4.2 Essential Dimensionality for Four 28-item General Tests with 3 Randomly Selected AT1.

	Target Test #AT1/#Total	AT1	Stout's T	P-value	Refined T	P-value
			.37	.35	.44	.33
	Test 1 (3/28)	Random	-.36	.64	-.56	.71
			.60	.27	.85	.20
			-.62	.73	-.72	.77
			.24	.40	.31	.38
	Test 2 (3/28)	Random	.38	.35	.39	.35
			.78	.22	1.00	.15
U.S.			-.99	.83	-1.27	.90
			1.51	.07	2.02	.02*
	Test 3 (3/28)	Random	.36	.36	.49	.31
			-.52	.70	-.66	.75
			-.66	.75	-.82	.79
			.83	.20	1.14	.13
	Test 4 (3/28)	Random	1.28	.10	1.64	.05
			-.03	.51	.00	.50
			-.93	.83	-1.21	.89

Table 4.3 shows the effects of reducing AT1 items from 12 to 8 and then to 4 on the essential dimensionality estimates for the original general tests 1 to 4. According to the results, three of four tests were multidimensional when the number randomly selected of AT1 items equaled 12. When the number of AT1 was 8 or 4, none of the tests are multidimensional except test 1 and 2 with 4 AT1 items. Thus, these results suggest that reducing the number of AT1 items may increase the possibility of inacurately concluding that a test is unidimensional. Taken together, the analyses in Table 4.2 and 4.3 do suggest that using either a smaller number of total items (Table 4.2) or a smaller number of AT1 items (Table 4.3) does increase the chance that a multidimensional test (i.e., mathematics genera achievement) will be identified as unidimensional.

Unfortunately, the analyses reported in Table 4.2 and 4.3 also are confounded because the selection of AT1 items was random. Prior analyses (e.g., Table 2 and Table 3) clearly suggest that a conclusion supporting unidimensionality depends on the nature of the arbitrarily selected AT1 items. It is intuitively reasonable that selecting AT1 items at random will result in stronger support for unidimensionality because the AT1 standard is itself more heterogeneous. To test this hypothesis, an additional analysis, analogous to that reported in Table 4.2, is shown in Table 4.4. In this analysis, three AT1 items within four mathematics subcontents --- decimal fractions, ratios, equations and estimations were selected to be homogeneous AT1 items (as recommended by Stout).

Table 4.4 shows that only four out of a total of sixteen trials of dimensionality assessment with homogeneous AT1 items demonstrated significant results. In other words, increasing the homogeneity of AT1 items does not produce uniformly significant essential dimensionality estimates. However, the four 28-item general tests were all flagged as multidimensional when decimal fractions items were the AT1 items. This finding, somewhat, suggested that the actual characteristics of AT1 items has a stronger influence on the essential dimensionality estimates than the number of AT1 items.

Table 4.3. Essential Dimensionality for Four SIMS General Tests with Three Sizes of AT1

Target Test #Total	AT1	Stout's T	P-value	Refined T	P-value
Test 1 (75)	12	-.03	.51	.00	.50
	8	-.93	.83	-1.21	.89
	4	1.83	.03*	2.01	.03*
Test 2 (75)	12	3.33	.00**	3.96	.00**
	8	.11	.45	.05	.48
	4	2.36	.01*	2.96	.01*
Test 3 (75)	12	2.05	.02*	2.34	.01*
	8	.29	.38	.25	.40
	4	.45	.33	.64	.26
Test 4 (75)	12	2.83	.00**	3.37	.00**
	8	.08	.46	.16	.42
	4	.74	.23	.94	.17

Table 4.4 Essential Dimensionality Statistics for Four 28-item General Tests with 3 Homogeneous AT1 Items.

Target Test #AT1/#Total	AT1	Stout's T	P-value	Refined T	P-value
Test 1 (3/28)	Deci. Frac.	2.14	.02*	2.72	.00**
	Ratios	.53	.30	.72	.23
	Equations	-.51	.70	-.72	.76
	Estimations	.01	.50	.05	.48
Test 2 (3/28)	Deci. Frac.	1.79	.04*	2.25	.01*
	Ratios	.53	.30	.64	.26
	Equations	.83	.20	1.09	.14
	Estimations	.82	.21	1.10	.14
Test 3 (3/28)	Deci. Frac.	1.63	.05*	2.09	.02*
	Ratios	-.68	.75	-.82	.79
	Equations	.69	.25	.88	.19
	Estimations	.14	.44	.23	.41
Test 4 (3/28)	Deci. Frac.	1.28	.10	1.64	.05*
	Ratios	.92	.18	1.18	.12
	Equations	.09	.46	.11	.46
	Estimations	.72	.24	1.00	.16

Robustness of DIMTEST.

To investigate the robustness of the essential dimensionality statistics, the relationship between the degree of the essential dimensionality of a test and the existence of item invariance property in the test was examined. Two levels of tests were used in this analysis. Test 1.1 and 1.2, arithmetic and algebra, were treated as essentially unidimensional tests based on results of Wang and Hocevar (1994), while the general mathematics test 1 was treated as multidimensional. Item parameters were estimated using a two-parameter logistic model (2-PL) with Bayes estimate procedure, which assumes a prior normal distribution of ability.

Studies indicated that violating the unidimensionality assumption produces a substantial lack of item parameter invariance (Ackerman, 1991; Oshima & Miller, 1990). Because Stout (1990) suggested that the essential dimensionality assumption is a valid substitute for IRT unidimensionality assumption, the item invariance property should fit essentially unidimensional tests better than multidimensional tests. In other words, the item invariance property is assumed to fit tests 1.1 and 1.2 significantly better than general test 1.

Table 5.1 presents the results of the IRT item invariance examination for the essentially unidimensional arithmetic test. An item was detected as lacking the item invariance property when either one of the three invariance measures was significant. It was unexpectedly found that 7 out of a total of 28 arithmetic items presented a lack of item invariance even when the test was essentially dimensional. This result provokes the need to reconsider the equivalence between the essential dimensionality assumption and the IRT unidimensionality assumption and the appropriateness of using the essential dimensionality assumption as a substitute for the IRT unidimensionality assumption.

Statistically, the sensitivity of Raju's exact unsigned measure is much stronger than the other two measures, and this may introduce some spurious detection. Only three items, furthermore, were identified by all three statistics as violating the item invariance property. The correlation between Lord's chi-square and Raju's signed area measure is higher than the other two possible pairings.

Table 5.2 presents a similar analysis on the essentially unidimensional algebra test 1.2. The invariance property does not hold for 4 items out of a total of 14 algebra items which were calibrated on an essentially unidimensional "algebra" scale. In three of the four cases, all three statistics uniformly indicated a lack of invariance. These results indicate that Stout's essential dimensionality is not a sufficient condition for the existence of the invariance property of the item parameters. However, this criticism is somewhat qualified in that only six of forty-two items were uniformly identified as lacking invariance by all three invariance indices.

Table 5.1. Item Parameter Estimates a and b and Three Item Invariance Indices for Esentially Unidimensional Arithmetic Test 1.1

	a		b							
Item	S1	S2	S1	S2	χ^2	p	ESA	p	EUA	p
ys076	1.41	-0.27	1.15	-0.36	2.90	0.23	-0.92	0.35	-1.66	0.09
ys075	0.86	0.90	1.00	1.13	8.26	0.01*	1.36	0.17	2.51	0.01*
ys003	0.92	0.27	0.93	0.64	9.40	0.00*	2.80	0.00*	2.80	0.00*
ys043	1.22	-0.53	1.11	-0.35	4.43	0.10	1.72	0.08	-2.03	0.04*
ys045	0.80	1.18	0.65	1.52	2.02	0.36	1.35	0.17	-1.36	0.17
ys109	1.86	-1.05	1.62	-0.95	4.80	0.09	1.03	0.29	-2.13	0.03*
ys005	1.37	-0.28	1.45	-0.17	1.69	0.42	1.24	0.21	1.28	0.19
ys140	1.07	-0.63	0.70	-0.63	9.60	0.00*	0.02	0.97	-2.55	0.01*
ys189	1.34	-0.48	1.23	-0.37	2.28	0.31	1.14	0.25	-1.40	0.15
ys079	2.18	-0.48	2.28	-0.40	1.55	0.45	1.23	0.21	1.24	0.21
ys181	1.51	0.17	1.53	0.26	1.22	0.54	1.03	0.30	1.03	0.29
ys190	1.01	-0.26	1.01	0.09	10.00	0.00*	3.15	0.00*	-3.15	0.00*
ys008	1.28	0.60	1.31	0.50	0.74	0.68	-0.83	0.40	0.82	0.41
ys179	1.20	-0.08	1.11	-0.04	0.48	0.78	0.44	0.65	-0.65	0.51
ys009	1.28	0.49	1.30	0.64	2.67	0.26	1.38	0.16	1.38	0.16
ys046	1.27	0.02	1.35	-0.03	0.53	0.76	-0.63	0.52	0.66	0.50
ys042	1.16	-1.36	1.03	-1.32	1.77	0.41	0.19	0.84	-0.84	0.39
ys074	1.42	-0.77	1.09	-0.80	4.22	0.12	-0.23	0.81	-1.85	0.06
ys106	1.64	-0.68	1.69	-0.59	1.01	0.60	1.00	0.31	0.99	0.32
ys185	0.70	1.91	0.71	2.06	1.10	0.57	0.45	0.64	0.47	0.63
ys187	0.96	0.61	1.04	0.72	1.96	0.37	0.81	0.41	1.22	0.21
ys077	0.68	-0.72	0.54	-0.56	4.28	0.11	0.81	0.41	-1.36	0.17
ys108	1.34	-0.65	1.30	-0.38	9.27	0.00*	2.77	0.00*	-2.77	0.00*
ys142	1.10	1.85	1.02	1.91	0.37	0.82	0.24	0.80	-0.44	0.65
ys191	1.00	-0.77	0.85	-0.86	1.15	0.56	-0.55	0.58	-1.04	0.29
ys192	1.20	-0.02	1.12	0.13	2.51	0.28	1.55	0.12	-1.50	0.13
ys048	1.24	0.24	1.42	0.21	1.19	0.54	-0.32	0.74	1.08	0.27
ys011	1.02	-0.30	1.14	-0.12	3.34	0.18	1.70	0.08	1.59	0.11

Note. S1 and S2 denote replications 1 and 2 which consist of approximately half of the original sample size.
* $p<.05$.

Table 5.2. Item Parameter Estimates a and b and Three Item Invariance Indices for Essentially Unidimensional Algebra Test 1.2.

	a		b							
Item	S1	S2	S1	S2	χ^2	p	ESA	p	EUA	p
ys012	1.95	-0.26	1.59	-0.34	2.97	0.22	-0.88	0.37	-1.67	0.09
ys013	1.29	0.11	1.21	0.28	2.29	0.31	1.50	0.13	-1.47	0.14
ys149	1.22	-0.37	1.44	-0.22	3.24	0.19	1.48	0.13	1.61	0.10
ys151	1.00	1.51	0.88	1.78	1.26	0.53	1.12	0.26	-1.06	0.28
ys017	1.29	-1.01	1.32	-0.72	6.69	0.03*	2.27	0.02*	2.27	0.02*
ys086	1.12	0.01	1.13	0.05	0.10	0.94	0.31	0.75	0.31	0.75
ys196	1.02	-0.39	1.03	-0.29	0.65	0.72	0.79	0.42	0.79	0.42
ys019	0.76	0.81	0.71	0.77	0.39	0.81	-0.20	0.83	-0.44	0.65
ys014	1.25	-0.27	1.12	-0.01	6.13	0.04*	2.35	0.01*	-2.42	0.01*
ys084	0.67	1.23	0.49	1.61	2.60	0.27	1.17	0.24	-1.49	0.13
ys195	1.66	0.03	1.68	0.12	0.94	0.62	0.92	0.35	0.92	0.35
ys115	2.30	0.30	1.75	0.40	3.77	0.15	1.03	0.30	-2.01	0.04*
ys053	1.76	-0.39	1.69	-0.16	6.40	0.04*	2.51	0.01*	-2.51	0.01*
ys087	0.54	2.90	0.41	3.94	1.40	0.49	1.18	0.23	-1.13	0.25

Table 5.3 presents the examination of the item parameter invariance property for multidimensional Test 1. It was expected that the item invariance property would fit an essentially unidimensional test better than a multidimensional test. Therefore, all items violating the invariance property in the essentially unidimensional test should not fit the invariance property in a multidimensional situation. Unfortunately, the expected result did not occur. In this table, it was found that only one-half of the twelve target items, which violated the invariance property in the two earlier unidimensional calibrations were identified as violating the invariance property in the multidimensional calibration. Most of the remaining six target items show moderately good fit to the item invariance property. In addition, the proportion of the items violating the item parameter invariance property in the multidimensional test (i.e., 23/75 or approximately .31) is not dramatically higher than its unidimensional arithmetic test (i.e., 7/28 or approximately .25) and algebra test (i.e., 4/14 or approximately .29). This result indicates that it may not be valid to use the essential dimensionality of a test as an index to determine the appropriateness of using unidimensional item calibration for a test.

For the purpose of exploring the appropriate level of SIMS mathematic content in which the item invariance property holds, the previous essentially unidimensional tests (i.e., arithmetic or algebra tests) were further split into three subscales. In this analysis three tests, common fractions, decimal fractions and ratio were selected from tests 3.1, 2.1 and 4.1, respectively. Items within the same arithmetic subarea were calibrated on a "unidimensional" arithmetic subscale. The existence of the item parameter invariance property for these calibrations was examined and the results are presented in Tables 6.1 to 6.3.

Table 6.1 shows that the invariance property of the item parameter is perfect at this level of item calibrations; that is, all six common fraction items in test C fit the item invariance property. The correlations between the three invariance indices, still, are low. The results of the existence of the invariance of the item parameters for the seven decimal fractions items in test 2.1 are shown in Table 6.2. The invariance property of the item parameters again fits this level of item calibration perfectly. A similar results was found for the 11 ratio items in test 4.1 and is displayed in Table 6.3.

In conclusion, the results in Tables 5.1 and 5.2 show that many essentially unidimensional items were detected to lack item invariance. Table 5.3 displays the invariance property of multidimensional items in test 1 and shows that multidimensional items in test 1 fit the invariance property almost as well as the essentially unidimensional tests. Both results indicate that Stout's essential unidimensionality assumption may not be a sufficient condition for the existence of the invariance property of item parameters.

Table 5.3. Item Parameter Estimates a and b and Three Item Invariance Indices for Essentially Multidimensional Test 1.

	a		b							
Item	S1	S2	S1	S2	χ^2	p	ESA	p	EUA	p
ys100	1.22	0.27	1.21	0.46	4.08	0.12	1.94	0.05	-1.94	0.05
ys189	1.36	-0.28	1.18	-0.36	1.82	0.40	-0.88	0.37	-1.32	0.18
ys151	0.92	1.54	0.83	1.82	1.34	0.51	1.14	0.25	-1.06	0.28
ys076	0.95	0.81	1.02	1.10	7.86	0.01*	1.89	0.05	2.09	0.03*
ys165	1.15	0.09	0.99	0.49	12.37	0.00*	3.51	0.00*	-3.29	0.00*
ys167	0.66	-0.37	0.88	-0.12	5.27	0.07	1.69	0.08	1.96	0.04*
ys031	1.33	-0.74	1.34	-0.47	8.46	0.01*	2.72	0.00*	2.72	0.00*
ys069	1.91	-1.58	1.73	-1.56	1.26	0.53	0.08	0.93	-0.75	0.45
ys038	1.24	0.05	1.24	0.23	3.63	0.16	1.90	0.05	1.90	0.05
ys168	0.48	4.72	0.42	5.56	0.47	0.78	0.61	0.53	-0.59	0.54
ys175	1.23	0.83	1.28	0.87	0.47	0.78	0.29	0.76	0.62	0.53
ys079	0.98	0.24	0.96	0.61	9.79	0.00*	2.94	0.00*	-2.94	0.00*
ys017	1.54	-0.92	1.50	-0.67	9.04	0.01*	2.55	0.01*	-2.55	0.01*
ys181	1.44	-0.49	1.29	-0.33	4.86	0.08	1.74	0.08	-2.10	0.03*
ys045	0.77	1.19	0.70	1.40	0.86	0.64	0.92	0.35	-0.80	0.41
ys012	1.71	-0.27	1.56	-0.33	1.29	0.52	-0.88	0.37	-1.11	0.26
ys121	1.03	-0.29	1.03	-0.19	0.79	0.67	0.88	0.37	0.88	0.37
ys086	1.12	-0.00	1.08	0.03	0.26	0.87	0.45	0.64	-0.47	0.63
ys023	0.78	1.21	0.78	1.21	0.00	0.99	-0.00	0.99	0.00	0.99
ys109	2.01	-1.00	1.74	-0.92	4.96	0.08	0.92	0.35	-2.13	0.03*
ys127	0.77	0.91	0.91	1.03	4.47	0.10	0.66	0.50	1.43	0.15
ys122	1.62	-0.28	1.38	-0.13	6.15	0.04*	1.93	0.05	-2.33	0.01*
ys103	1.45	-0.04	1.49	0.26	14.00	0.00*	3.56	0.00*	3.56	0.00*
ys190	1.54	-0.27	1.54	-0.18	1.40	0.49	1.17	0.23	-1.17	0.23
ys013	1.66	0.06	1.34	0.24	7.05	0.02*	2.12	0.03*	-2.51	0.01*
ys005	1.00	-0.66	0.77	-0.61	6.60	0.03*	0.30	0.75	-2.07	0.03*
ys149	1.32	-0.37	1.40	-0.22	3.10	0.21	1.72	0.08	1.74	0.08
ys075	1.49	-0.46	1.19	-0.38	4.97	0.08	0.82	0.40	-2.03	0.04*
ys068	1.69	-0.99	1.76	-0.82	3.56	0.16	1.77	0.07	1.77	0.07
ys019	0.86	0.70	0.79	0.68	0.56	0.75	-0.11	0.91	-0.59	0.55
ys003	2.03	-0.48	1.97	-0.42	0.97	0.61	0.89	0.37	-0.93	0.34
ys140	1.47	0.15	1.38	0.26	1.49	0.47	1.18	0.23	-1.18	0.23
ys008	1.04	-0.26	1.00	0.08	10.24	0.00*	3.14	0.00*	-3.14	0.00*
ys179	1.19	0.60	1.19	0.52	0.66	0.71	-0.72	0.46	0.72	0.46
ys196	1.08	-0.39	1.07	-0.29	0.90	0.63	0.90	0.36	-0.90	0.36
ys009	1.20	-0.09	1.08	-0.05	0.82	0.66	0.44	0.65	-0.87	0.38
ys043	1.43	0.43	1.33	0.61	3.21	0.20	1.78	0.07	-1.71	0.08
ys046	1.18	0.01	1.27	-0.04	0.72	0.69	-0.66	0.50	0.77	0.43
ys028	1.13	-0.04	0.97	-0.08	1.41	0.49	-0.33	0.74	-1.17	0.24
ys156	0.94	0.12	1.01	0.33	4.40	0.11	1.81	0.06	2.09	0.03*
ys042	1.27	-1.28	1.08	-1.28	2.36	0.30	-0.03	0.96	-1.06	0.28
ys030	1.16	-1.08	0.84	-1.04	9.26	0.00*	0.23	0.81	-2.15	0.03*
ys185	1.47	-0.75	1.12	-0.79	5.03	0.08	-0.34	0.73	-2.06	0.03*
ys087	0.61	2.56	0.42	3.79	2.91	0.23	1.60	0.10	-1.58	0.11
ys195	1.65	0.01	1.81	0.10	2.43	0.29	1.16	0.24	1.52	0.12
ys025	0.71	2.37	0.54	3.07	2.23	0.32	1.30	0.19	-1.37	0.16
ys101	0.95	-0.87	0.80	-0.87	1.82	0.40	-0.01	0.98	-1.09	0.27
ys171	0.27	4.76	0.21	6.47	0.78	0.67	0.86	0.38	0.80	0.42
ys072	0.84	0.79	0.85	0.77	0.01	0.99	-0.11	0.90	0.10	0.91
ys058	0.52	1.40	0.55	1.66	2.59	0.27	0.73	0.46	1.39	0.16
ys132	0.75	-2.30	1.08	-1.79	3.77	0.15	1.41	0.15	1.62	0.10
ys097	0.82	-0.23	0.73	-0.00	3.33	0.18	1.65	0.09	-1.37	0.17

Table 5.3. (Continued) Item Parameter Estimates a and b and Three Item Invariance Indices for Essentially Multidimensional Test 1.

	a		b							
Item	S1	S2	S1	S2	χ^2	p	ESA	p	EUA	p
ys187	1.59	-0.68	1.56	-0.61	0.75	0.68	0.73	0.46	-0.77	0.43
ys176	0.36	3.03	0.38	3.38	2.13	0.34	0.41	0.67	0.50	0.61
ys159	0.73	-0.65	0.67	-0.65	0.38	0.82	0.00	0.99	-0.53	0.59
ys142	0.69	1.90	0.80	1.86	1.98	0.37	-0.13	0.89	0.83	0.40
ys074	1.01	0.56	1.10	0.67	2.17	0.33	0.89	0.37	1.35	0.17
ys077	0.74	-0.69	0.59	-0.53	4.63	0.09	0.87	0.38	-1.47	0.14
ys066	1.44	-0.47	1.31	-0.20	11.27	0.00*	3.13	0.00*	-3.28	0.00*
ys070	0.67	0.76	0.71	1.04	4.16	0.12	1.35	0.17	1.88	0.05
ys108	1.41	-0.63	1.30	-0.39	9.61	0.00*	2.61	0.00*	-2.82	0.00*
ys033	1.97	-1.28	2.05	-1.14	2.27	0.31	1.32	0.18	1.32	0.18
ys191	1.15	1.77	1.06	1.84	0.42	0.80	0.33	0.73	-0.51	0.60
ys048	1.13	-0.72	0.91	-0.82	2.47	0.29	-0.75	0.44	-1.50	0.13
ys192	1.20	-0.03	1.16	0.11	2.36	0.30	1.53	0.12	-1.53	0.12
ys115	1.82	0.31	1.51	0.41	3.37	0.18	1.18	0.23	-1.85	0.06
ys022	1.15	-0.39	1.08	-0.33	0.85	0.65	0.65	0.50	-0.77	0.43
ys053	1.74	-0.39	1.59	-0.16	10.09	0.00*	3.02	0.00*	-3.12	0.00*
ys194	0.62	1.32	0.46	1.87	2.48	0.28	1.48	0.13	-1.47	0.14
ys011	1.44	0.19	1.59	0.17	0.82	0.66	-0.22	0.82	0.90	0.36
ys057	0.85	0.01	0.72	0.38	6.72	0.03*	2.55	0.01*	-2.03	0.04*
ys106	1.12	-0.30	1.21	-0.13	3.04	0.21	1.68	0.09	1.65	0.09
ys084	0.77	1.06	0.59	1.34	2.50	0.28	1.13	0.25	-1.50	0.13
ys029	0.89	1.52	0.54	2.34	7.95	0.01*	2.14	0.03*	-2.40	0.01*
ys014	1.40	-0.27	1.28	-0.03	8.07	0.01*	2.76	0.00*	-2.83	0.00*

Note. Items underlined denote violating the invariance property in essentially unidimensional test.

The invariance property exists when essentially unidimensional arithmetic tests are split into three subtests in which the smallest contains only six items (Table 6.1). This result contradicts the finding of prior studies which concluded that item parameters are more stable in a longer test than a shorter test (Shepard, Camilli & William, 1985, Subkoviak, Mack, Ironson & Craig, 1984). A possible implication for this result is that unidimensionality has a stronger influence than the number of items on the stability of item parameter estimation. A follow-up study is needed to make this issue clear

Table 6.1. Item Parameter Estimates a and b and Three Item Invariance Indices for Common Fractions in Test 3.1.

	a		b							
Item	S1	S2	S1	S2	χ^2	p	ESA	p	EUA	p
ys003	3.18	2.61	-.31	-.33	.84	.67	-.36	.72	-.89	.37
ys004	1.07	1.19	-1.50	-1.32	.24	.88	.49	.63	.44	.66
ys043	1.26	1.25	.72	.69	.09	.95	-.19	.85	-.30	.76
ys044	.76	.75	.58	.20	2.01	.37	-1.25	.21	-1.25	.21
ys185	1.24	.95	-.88	-.86	1.51	.47	.09	.93	-.97	.33
ys186	.52	.69	-2.02	-2.00	1.72	.42	.03	.98	.72	.47

Table 6.2. Item Parameter Estimates a and b and Three Item Invariance Indices for Decimal Fractions in Test 2.1.

	a		b							
Item	S1	S2	S1	S2	χ^2	p	ESA	p	EUA	p
ys005	1.01	1.04	-.53	-.48	.22	.90	.47	.64	.37	.71
ys006	.93	1.08	.56	.59	.48	.79	.09	.93	.54	.59
ys007	1.38	1.32	-.44	-.41	.05	.98	.15	.88	-.16	.87
ys045	.79	.81	1.22	1.15	.15	.93	-.31	.76	.27	.79
ys109	1.67	1.70	-.90	-.95	.46	.79	-.46	.65	.46	.65
ys140	1.31	1.30	.43	.45	.03	.99	.17	.86	-.15	.86
ys141	.76	.98	.11	-.01	.90	.64	-.55	.58	.85	.39

Table 6.3 Item Parameter Estimates a and b and Three Item Invariance Indices for Ratio Items in Test 4.1.

	a		b							
Item	S1	S2	S1	S2	χ^2	p	ESA	p	EUA	p
ys008	.95	.99	.05	.06	.14	.93	-.29	.78	.33	.74
ys009	1.33	1.32	-.11	-.09	.05	.97	.23	.82	-.22	.82
ys046	1.31	1.46	-.03	.02	1.07	.59	.47	.64	.91	.36
ys047	1.31	.96	-.26	-.24	1.74	.42	.03	.98	-1.27	.20
ys079	.65	.95	1.79	1.94	4.92	.09	.26	.79	1.38	.17
ys110	1.00	1.00	-.02	.05	.07	.96	.27	.78	-.27	.78
ys142	.69	.68	2.09	1.81	.79	.67	-.48	.63	-.48	.63
ys143	.78	.55	-.42	-.40	1.32	.52	.06	.95	-1.04	.30
ys190	1.23	1.12	-.26	-.25	.45	.80	.05	.96	-.67	.50
ys191	1.06	1.06	2.10	1.73	1.84	.40	.88	.38	.88	.38
ys192	1.33	1.19	.04	.05	.21	.90	.02	.99	-.45	.65

Conclusions and Discussion

The reliability and validity of Stout's essential dimensionality statistics were examined in this study. The stability of two essential dimensionality measures was found to be low for some tests across ten random samples. The cause of this difference is unclear because the effect of the interaction between respondents and items are confounded with the effect of the reliability of Stout's measures. If we can declare that the cognitive ability space is the same across groups of random samples, we can conclude that Stout's two essential dimensionality measures are somewhat unreliable. The essential dimensionality results for the four tests across four AT1 assignments was also different which indicates that the essential dimensionality estimate for a test is related to the characteristics of the AT1 items.

Two substantive findings in the first analysis are first, four algebraic tests tend to be more consistently identified as essentially unidimensional than their arithmetic counterparts, and second, Stout's original essential dimensionality measure is more consistent than the refined statistic which was proposed by Nandakumar (1993).

In the second section of analysis, the effect of reducing the number of examinees and test items was analyzed. It was found that reducing sample size does not provide consistent improvement on the

degree of the essential unidimensionality. Small sample size does however cause a fatal problem in running DIMTEST. The degree of essential unidimensionality tended to increase when the number of test items and number of AT1 items decreased. But, test 2 was flagged as multidimensional even when the number of AT1 items was reduced to 4. Therefore, reducing the number of test items and AT1 items does not assure unidimensionality. The characteristics of AT1 items likely has more influence on the essential dimensionality estimates.

The validity of replacing the IRT unidimens.. ..ality assumption by the essential dimensionality assumption was assessed at the last section of the analysis using the invariance of item parameters as evidence. It was universally found that the relationships between the existence of the item invariance property and the essentially unidimensional item calibrations (i.e., arithmetic and algebra scale) are low across test forms and mathematic areas. The degree of the fit of the item invariance property for two "essentially unidimensional" tests (75% and 71%) are approximately the same as the "essentially multidimensional" test 1 (69%). A possible interpretation is that since the degree of the essential dimensionality of a test is related to the characteristics of the AT1 items, the degree of dimensionality for a test can not be meaningfully determined unless "appropriate" AT1 items are determined. Therefore, a further study on the criteria for AT1 items is needed to enhance the validity of replacing the IRT unidimensionality assumption by the essential dimensionality assumption.

For the purpose of determining on which mathematical level the item invariance property exists, four essentially unidimensional arithmetic tests were further split into three subtests. Three logically constructed subtests across three test forms (i.e., common fractions in test 3.1, decimal fractions in test 2.1 and ratios in test 4.1) fit the item parameter invariance property consistently well.

References

Ackerman, T. A. (1991). The use of unidimensional item parameter estimstes of multidimensional items in adaptive testing. Applied Psychological Measurement, 15, 13-24.

Chang, L. C., & Ruzicka, J. (1985). Second International Mathematics Study: United States Technical report I. University of Oklahoma, National Research Coordinator.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Lawrence Erlbaum Associates.

Nandakumar, R (1993). Assessing essential unidimensionality of real data. Applied Psychological Measurement, 17, 29-38.

Oshima, T. C., & Miller, M. D. (1990). Multidimensionality and IRT-based item invariance indexes: The effect of between-group variation in trait correlation. Journal of Educational Measurement, 27, 273-283.

Raju, N. S. (1988). The area between two item characteristic curves. Psychometrika, 53, 495-502.

Shepard, L. A., Camilli, G., & Williams, D. M. (1985). Validity of approximation techniques for detecting item bias. Journal of Educational Measurement, 22, 77-105.

Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. Psychometrika, 52, 589-617.

Stout, W. (1990). A new item response theory modeling approach with applications to unidimensionality assessment and ability estimation. Psychometrika, 55, 293-325.

Stout, W., Douglas, J., Junker, B., Roussos, L. (1992). DIMTEST [Computer Program]. Urbana-Champaign, IL: University of Illinois.

Subkoviak, M. J., Mack, J. S., Ironson, G. H., & Craig, R. D. (1984). Empirical comparison of selected item bias detection procedures with bias manipulation. Journal of Educational Measurement, 21, 49-58.

Traub, R. E. (1983) A prior considerations in choosing a item response model. In R. K. Hambleton,. (Ed.). Applications of item response theory (pp. 57-70). Vancouver, BC: Educatior. l Research Institute of British Columbia.

Wang, Y. L., & Hocevar, D. (1994, April). Effects of mathematics test content specificity on essential dimensionality in the U.S. and Japan data. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.